Abelson (1995).

Statistics as Principled Argument

1 Making Claims With Statistics

MISUNDERSTANDINGS OF STATISTICS

The field of statistics is misunderstood by students and nonstudents alike. The general public distrusts statistics because media manipulators often attempt to gull them with misleading statistical claims. Incumbent politicians, for example, quote upbeat economic statistics, whereas their challengers cite evidence of wrack and ruin. Advertisers promote pills by citing the proportion of doctors who supposedly recommend them, or the average time they take to enter the bloodstream. The public suspects that in the interest of making particular points, propagandists can use any numbers they like in any fashion they please.

Suspicion of false advertising is fair enough, but to blame the problem on statistics is unreasonable. When people lie with words (which they do quite often), we do not take it out on the English language. Yes, you may say, but the public can more readily detect false words than deceitful statistics. Maybe true, maybe not, I reply, but when statistical analysis is carried out responsibly, blanket public skepticism undermines its potentially useful application. Rather than mindlessly trashing any and all statements with numbers in them, a more mature response is to learn enough about statistics to distinguish honest, useful conclusions from skullduggery or foolishness.

It is a hopeful sign that a considerable number of college and university students take courses in statistics. Unfortunately, the typical statistics course does not deal very well, if at all, with the argumentative, give-and-take nature of statistical claims. As a consequence, students tend to develop their own characteristic misperceptions of statistics. They seek certainty and exactitude, and emphasize calculations rather than points to be drawn from statistical analysis. They tend to state statistical conclusions mechanically, avoiding imaginative rhetoric (lest they be accused of manipulativeness).

It is the aim of this book to locate the field of statistics with respect to rhetoric and narrative. My central theme is that good statistics involves principled argument that conveys an interesting and credible

point. Some subjectivity in statistical presentations is unavoidable, as acknowledged even by the rather stuffy developers of statistical hypothesis testing. Egon Pearson (1962), for example, wrote retrospectively of his work with Jerzy Neyman, "We left in our mathematical model a gap for the exercise of a more intuitive process of personal judgment" (p. 395). Meanwhile, Sir Ronald Fisher (1955) accused Neyman and Pearson of making overmechanical recommendations, himself emphasizing experimentation as a continuing process requiring a community of free minds making their own decisions on the basis of shared information.

Somewhere along the line in the teaching of statistics in the social sciences, the importance of good judgment got lost amidst the minutiae of null hypothesis testing. It is all right, indeed essential, to argue flexibly and in detail for a particular case when you use statistics. Data analysis should not be pointlessly formal. It should make an interesting claim; it should tell a story that an informed audience will care about, and it should do so by intelligent interpretation of appropriate evidence from empirical measurements or observations.[1]

CLAIMS MADE WITH STATISTICS: COMPARISON AND EXPLANATION

How are claims developed in statistical tales? For most of this book, we treat statistics in connection with systematic research programs, but to begin, let us discuss the case in which purportedly newsworthy statistical "facts" are picked up by roving reporters and presented in the media.

Stand-Alone Statistics

Many of these statistics are isolated, stand-alone figures such as: "The average life expectancy of famous orchestral conductors is 73.4 years" (Atlas, 1978), or "adults who watched television 3–4 hours a day had nearly *double* the prevalence of high cholesterol as those who watched less than one hour a day" (Tucker & Bagwell, 1992), or "...college-educated women who are still single at the age of thirty-five have only a 5 percent chance of ever getting married" ("Too Late," 1986; discussed by Cherlin, 1990; and Maier, 1991). The point of the life-expectancy statistic was supposedly that conducting an orchestra is so fulfilling that it lengthens life. The cholesterol story was somewhat puzzling, but the implication was that increased junk food consumption accompanied heavy TV watching. The marriage statistic was based on shaky projections of future trends, and could be variously explained or dismissed, depending on who was doing the explaining or dismissing.

A problem in making a claim with an isolated number is that the audience may have no context within which to assess the meaning of the figure and the assertion containing it. How unusual is it to live until age 73.4? Does "nearly double" mean I shouldn't watch TV? If one can't answer such questions, then a natural reaction to this type of numerical pronouncement would be, "So what?"

The Importance of Comparison

In the example about women and percentage marrying, a background context is readily available, and most people would regard 5% as a startlingly low marriage rate compared to the general average (or compared to what was true 50 years ago). The idea of *comparison* is crucial. To make a point that is at all meaningful, statistical presentations must refer to differences between observation and expectation, or differences among observations. Observed differences lead to why questions, which in turn trigger a search for explanatory factors. Thus, the big difference between the 5% future marriage rate for 35-year-old, college-educated single women and one's impression that some 80% or 90% of women in general will marry, evokes the question, "I wonder why that is? Is it career patterns, the lack of appeal of marriage, or a shortage of eligible men?... Or maybe the 5% figure is based on a faulty statistical procedure." Such candidate explanations motivate the investigators (or their critics) to a reanalysis of existing evidence and assumptions, or the collection of new data, in order to choose a preferred explanation.

Apart from the standard statistical questions of why there is a difference between one summary statistic and another, or between the statistic and a baseline comparison figure, there occasionally arises a need to explain a lack of difference. When we expect a difference and don't find any, we may ask, "Why is there *not* a difference?" Galileo's fabled demonstration that heavy and light objects take the same time to fall a given distance is a case in point. The observed constancy stands in contrast with a strong intuition that a heavy object should fall faster, thus posing a puzzle requiring explanation.

[1]There is a different light in which some people view the field of statistics. Data gathering may be seen as the archival activity of assembling "facts," which at some later time may be used according to the needs of particular investigators or administrators. Historically, statistics began with the collection of tax and census records, and the term *statistics* derives from the description of *states* (Cowles, 1989). Prime modern examples of archives that can later be used for various research purposes are census data, and public opinion survey data banks such as the General Social Surveys (Davis & Smith, 1991). Resources like these are important, and I do not underestimate their value. Nevertheless, data banking is but the beginning of certain research enterprises, not their culmination. It is the theoretical and applied payoff of data analysis that engages my attention in this book.

Standards of Comparison

At the outset of the explanation process, there is a complication. Given a single statistic, many different observations or expectations may be used as standards of comparison; what is compared with what may have a substantial influence on the question asked and the answer given. Why questions are said to have a *focus.*[2] The longevity datum on famous orchestral conductors (Atlas, 1978) provides a good example. With what should the mean age at their deaths, 73.4 years, be compared? With orchestral *players*? With *nonfamous* conductors? With the general public?

All of the conductors studied were men, and almost all of them lived in the United States (though born in Europe). The author used the mean life expectancy of males in the U.S. population as the standard of comparison. This was 68.5 years at the time the study was done, so it appears that the conductors enjoyed about a 5-year extension of life—and indeed, the author of the study jumped to the conclusion that involvement in the activity of conducting *causes* longer life. Since the study appeared, others have seized upon it and even elaborated reasons for a causal connection (e.g., as health columnist Brody, 1991, wrote, "it is believed that arm exercise plays a role in the longevity of conductors" [p. B8]).

However, as Carroll (1979) pointed out in a critique of the study, there is a subtle flaw in life-expectancy comparisons: The calculation of average life expectancy includes infant deaths along with those of adults who survive for many years. Because no infant has ever conducted an orchestra, the data from infant mortalities should be excluded from the comparison standard. Well, then, what about teenagers? They also are much too young to take over a major orchestra, so their deaths should also be excluded from the general average. Carroll argued that an appropriate cutoff age for the comparison group is at least 32 years old, an estimate of the average age of appointment to a first orchestral conducting post. The mean life expectancy among U.S. males who have already reached the age of 32 is 72.0 years, so the relative advantage, if any, of being in the famous conductor category is much smaller than suggested by the previous, flawed comparison. One could continue to devise ever and ever more finely tuned comparison groups of nonconductors who are otherwise similar to conductors. Thoughtful attention to comparison standards (usually "control groups") can substantially reduce the occurrence of misleading statistical interpretations.

[2]A shift in the focus of question and answer is well illustrated by a joke beloved among 10-year-old children: "Why did the turkey cross the road?" ... "Because it was the chicken's day off." The reader who doesn't understand children's jokes can get the idea of focus by studying the effect of underlining different words in a why question. See Lehnert (1978).

Choosing Among Candidate Explanations

For any observed comparative difference, several possible candidate explanations may occur to the investigator (and to critics). In a given case, this set of explanations may include accounts varying widely in their substance and generality, ranging from a dismissal of the observed difference as a fluke or an artifactual triviality to claims that the observations support or undermine some broad theoretical position. In our orchestra conductors example, the set of candidate explanations includes at least the following: (a) The result arose fortuitously from the particular sample of conductors included; (b) the comparison standard is still flawed, as it does not account for subpopulations with shorter life spans who are also ineligible to become conductors (e.g., the chronically ill); and (c) conductors *do* live longer, because of some common genetic basis for longevity and extraordinary musical talent, health benefits from the activity of conducting (or from a larger class of activities that includes conducting), or health benefits from something associated with conducting, such as receiving adulation from others, or having a great deal of control over others.

It is the task of data analysis and statistical inference to help guide the choice among the candidate explanations. The chosen explanation becomes a *claim.* (If this term implies more force than appropriate, we may use the blander word *point.*) In the conductor example, it is risky to make a claim, because of a lack of relevant data that would help winnow the set of explanations. It would be helpful to have information on such matters as the life expectancy of well-known pianists, actors, professors, lawyers, and so forth; the life expectancy of eminent conductors who retire early (for reasons other than health); the life expectancy of siblings of famous conductors (ideally, twin siblings—but there would not be enough cases); and the comparative life expectancies of elderly people who stay active and those who are inactive (for reasons other than poor health).

Experimentalists would despair at the vagueness of specification of the needed evidence (how should one define "poor health," "active" "retire"), and the sinking feeling that there are just too many variables (some of them unknown) that might be associated with longevity. The experimental investigator would be in a much more comfortable position if he or she could isolate and *manipulate* the factors assumed to be relevant in one or more of the proposed causal accounts. An experimenter, as distinct from an observer, tries to *create* (or re-create) comparative differences rather than just to observe them passively.

Consider the possible explanation that orchestral conducting is so personally satisfying or otherwise beneficial that it extends life beyond the age at which the individual would have died in the absence of this activity. The standard experimental way to try to recreate such an effect

would be to assemble a group of potentially outstanding conductors, arrange for a random half of them to have prestigious orchestra posts whereas the other half have less involving career activities, and then collect longevity data on all of them. Of course this test would be absurdly impractical. I mention it because it suggests the possibility of conceptually similar experiments that might be feasible. For example, one could recruit a group of elderly people, provide a random half of them with social or physical activities, or social control, and monitor their subsequent feelings of well-being and state of health relative to that of the other half, who had received no intervention.[3] The bottom line for the conductors example, though, is that casual, one-shot tabulations of statistical observations will almost certainly be difficult to interpret. Therefore it is rhetorically weak to make claims based on them, and such claims deserve to be regarded with great skepticism. Well-justified explanations of comparative differences typically depend on well-controlled comparisons such as can be provided by careful experiments, and therefore we emphasize experimental data in this book. (Sometimes, one can also do well by the sophisticated collection of converging lines of evidence in field observations.) The quality of explanation improves dramatically when there are many interrelated data sets, some of them repeated demonstrations of the core result(s) or of closely related results, some of them ruling out alternative explanations, and yet others showing that when the explanatory factor is absent, the result(s) fail to appear.

Systematic Versus Chance Explanations

To understand the nature of statistical argument, we must consider what *types* of explanation qualify as answers to why questions. One characteristic type, the *chance* explanation, is expressed in statements such as, "These results could easily be due to chance," or "A random model adequately fits the data." Indeed, statistical inference is rare among scientific logics in being forced to deal with chance explanations as alternatives or additions to systematic explanations.

In the discussion to follow, we presume that data are generated by a single measurement procedure applied to a set of objects or events in a given domain. We suppose that the observations comprising the data set differ, some from others, and we ask why. A *systematic factor* is an influence that contributes an orderly relative advantage to particular subgroups of observations, for example, a longevity gain of a certain number of years by elderly people who stay active. A *chance factor* is an

[3]There is in fact a growing literature in the field of health psychology that speaks to precisely this idea (Langer & Rodin, 1976; Okun, Olding, & Cohn, 1990; Rodin, 1986).

influence that contributes haphazardly to each observation, with the amount of influence on any given observation being unspecifiable.

The Tendency to Exaggerate Systematic Factors

Inexperienced researchers and laypeople alike usually overestimate the influence of systematic factors relative to chance factors. As amateur everyday psychologists and would-be controllers of the world around us, we exaggerate our ability to predict the behavior of other people. We have difficulty thinking statistically about human beings.

Kunda and Nisbett (1986) showed that in matters of human *ability*, especially athletic ability, there is some degree of appreciation of inexplicable variations in performance from one occasion to the next. We understand, for example, that a tennis player might be on his game one day but flat the next, so that a sample of performances is necessary to make a reliable judgment of ability. Even so, the relative importance of chance influences is seriously underestimated in many athletic contexts. Abelson (1985) asked baseball-wise psychologists to consider whether or not a major league batter would get a hit in a given turn at bat, and to estimate the proportion of variance in this event explained by differences in the skill of different batters, as opposed to chance factors affecting the success of a given batter. The median estimate was around 25%, but the true answer is less than one half of 1%! In part this is due to the highly stingy properties of "explained variance" as a measure of relationship between two variables (Rosenthal & Rubin, 1979), but more interestingly, it is because we as baseball fans are prone to regard a .330 hitter as a hero who will almost always come through in the clutch, and the .260 hitter as a practically certain out when the game is on the line.

The underappreciation of chance variability extends to other domains. For events such as lottery drawings in which skill plays no objective role whatever, subjects under many conditions act as though some control can be exerted over the outcome (Langer, 1975). Kunda and Nisbett (1986) concluded that in matters of *personality*, inferences based on a single encounter are made with undue confidence, ignoring the possibility of situational influences that vary over time and place. We tend to feel, for example, that the person who is talkative on one occasion is a generally talkative person (the "fundamental attribution error," Ross, 1977).

The upshot of all this is a natural tendency to jump to systematic conclusions in preference to chance as an explanation. As researchers, we need principled data-handling procedures to protect us from inventing elaborate overinterpretations for data that could have been dominated by chance processes. We need to understand that even though statistical calculations carry an aura of numerical exactitude, debate

necessarily surrounds statistical conclusions, made as they are against a background of uncertainty. A major step in the winnowing of explanations for data is to make a judgment about the relative roles played by systematic and chance factors.

Inasmuch as chance is not well understood—even by those who have had a bit of statistical training—we introduce whimsical, hopefully memorable metaphors for the operation of chance factors (chap. 2).

LANGUAGE AND LIMITATIONS OF NULL HYPOTHESIS TESTS

A staple procedure used in psychological research to differentiate systematic from chance explanations is the significance test of a null hypothesis. Elementary statistics texts describe many varieties of them, but students often regard null hypothesis testing as counterintuitive, and many critics (e.g., Cohen, in press; Falk & Greenbaum, in press; Tukey, 1991) find much to fault in null hypothesis tests. It is worthwhile to set forth here the quirky logic of these tests, so that later on when we refer to their application, the reader will be well informed about their role in statistics, and the reasons for complaint about them.

Consider the simplest type of laboratory experiment, in which subjects are assigned at random to either an experimental group or a control group. Members of the two groups perform the identical experimental task, except for the additional manipulation of a single factor of interest in the experimental group—say, the receipt of prior information or training, or the administration of a drug. The experimenter wishes to test whether the experimental factor makes a systematic difference on some appropriate measure of task performance.

Presumably, performance measures on the task differ from individual to individual, and we ask a rhetorical why question. The systematic explanatory factor is the manipulation introduced by the experimenter. To say that this factor is systematic is to assume that on average it improves (or damages) task performances in the experimental group by some unknown amount over and above performances in the control group. We can try to estimate the magnitude of this systematic effect simply by calculating the difference between the mean performance scores of the two groups.

But there are also chance factors in this situation—things that add noise to individual measurements in an unknown way. We mention two categories here: sampling errors and measurement errors. Sampling errors arise from the "luck of the draw" in randomly assigning subjects to the two groups; the experimental group may contain a predominance of people with somewhat higher (or lower) task ability than members of the control group, thus introducing a mean difference that could be

mistaken for a systematic effect. Measurement errors refer to unknown and unrepeatable causes of variability in task performance over time, place, and circumstance. The laboratory room may be too warm when Subject 17 performs the task; Subject 42 may have a headache that day; and so on.

In qualitative terms, there are three possible accounts for the data arising from this experimental design: (a) The variability of task scores can be entirely explained by the systematic factor, (b) the variability of task scores can be entirely explained by chance factors (sampling and measurement errors), or (c) the variability requires explanation by both chance factors and the systematic factor.

The first and the second accounts are simpler, and parsimony would suggest that they be tested before falling back on the third account. Why tell a complicated story if a simpler story will do? The third account can be held in reserve if both of the first two accounts are inadequate. The first possibility, completely systematic data with no chance variability, would be immediately apparent in the data set: All the scores in the experimental group would be equal, and different from all the equal scores in the control group. This outcome may be approximated in the physical and biological sciences, where chance variability is typically very small. With psychological data, however, this outcome is quite rare—but if and when it occurs, statistical inference is not used (Skinner, 1963).

Setting aside these rare, errorless cases, we are left with the choice between the all-chance explanation, and the systematic-plus-chance explanation. We can tell if we need to invoke a systematic factor by first testing the all-chance explanation; if chance factors do not adequately account for the data, then the systematic factor is needed. This is in essence the justification for significance tests of the null hypothesis.

The Language of Null Hypothesis Testing

A null hypothesis test is a ritualized exercise of devil's advocacy. One assumes as a basis for argument that there is no systematic difference between the experimental and control scores—that except for errors of sampling and measurement the two groups' performances are indistinguishable. If (according to a formal procedure such as a t test) the data are not sharply inconsistent with this conception, then an all-chance explanation is tenable, so far as this one data set is concerned. This is often described as "accepting the null hypothesis." If, on the other hand, the data are inconsistent with the all-chance model, the null hypothesis is rejected, and the systematic-plus-chance model is preferred.

An important caveat here is that the standard terms, "accept" or "reject" the null hypothesis, are semantically too strong. Statistical tests are aids to (hopefully wise) judgment, not two-valued logical declara-

tions of truth or falsity. Besides, common sense tells us that the null hypothesis is virtually never (Cohen, 1990; Loftus, 1991) literally true to the last decimal place. It is thus odd to speak of accepting it. We often use other terms for this outcome, such as "retaining the null hypothesis" or "treating the null hypothesis as viable."[4] Similarly, rejection can be softened with alternative phrases like, "discrediting the null hypothesis."

In any case, the investigator wanting to show the influence of some experimental factor proceeds by discrediting the assumption that it doesn't matter. The backhandedness of this procedure reflects the fact that null hypothesis tests are motivated by rhetorical considerations. Suppose an experimental investigator announces that the data demonstrate—despite considerable variability from case to case—the systematic efficacy of a particular educational or medical intervention or the operation of a particular theoretical principle, but a critic counters that the data could easily have arisen from fortuitous sampling or measurement errors. Who wins this scientific debate? The critic does, unless the investigator can come up with a *counter*-counter, to the effect that the data are in fact quite unlikely to be explained entirely by chance factors. With such a rebuttal, the investigator discredits the null hypothesis (and therefore the critic will in practice usually be deterred from raising this argument in the first place).

Significance Tests Provide Very Limited Information

The answer to the simple question, "Is there some systematic difference between the experimental group and the control group?" is not usually electrifying. As mentioned earlier, there is virtually always some difference caused by sensible experimental manipulations. Indeed, the only examples where the exact satisfaction of the null hypothesis is worth considering occur when there is widespread disbelief that some strange phenomenon exists at all. For example, the null hypothesis is interesting when discrediting it implies that mental telepathy is possible, or that stimuli below the level of conscious awareness can have reliable effects on attitudes and behavior. The complementary case is also interesting, in which everybody believes beforehand that an effect must exist. For instance, virtually everyone who follows sports believes that there is such a thing as "streak shooting" in basketball, and it caused a considerable furor when Gilovich, Vallone, and Tversky (1985) argued from a set of sensitive statistical tests on sequences of shots that the null hypothesis of no streak shooting is tenable.

[4]A good way to think about what it means to retain a null hypothesis of no mean difference is that the analyst is insufficiently confident to assert which mean is larger (Tukey, 1991). See chapter 3, footnote 1.

Single Studies Are Not Definitive

Even in these rare cases, though, where the outcome of a simple significance test may have scientific (and possibly popular) news value, a single study never is so influential that it eliminates all argument. Replication of research findings is crucial. After all, if a result of a study is contrary to prior beliefs, the strongest holders of those prior beliefs will tend to marshall various criticisms of the study's methodology, come up with alternative interpretations of the results, and spark a possibly long-lasting debate.

Sometimes critics prove right in the long run, and sometimes they prove wrong. To take an example from the physical sciences, skepticism about the existence of "cold fusion" prevailed after a year or two of debate (Pool, 1988) over a claim of success. The opposite outcome of debate is illustrated by the reality of "subliminal perception"—meaningful stimulus registration without awareness—that after a period of skepticism has been widely accepted (Kihlstrom, 1987).

Debate about the existence of extrasensory perception (ESP) went on for years in an inconclusive, rather sterile fashion (Hyman, 1991; Utts, 1991). Argument had not progressed much beyond the "mere existence," null-hypothesis-testing stage to a more interesting, focused examination of the strength and generality (if any) of ESP, the conditions that encourage it, and the process by which it may operate. (Recently, the debate has become more usefully focused on the properties of a particular type of demonstration, the *Ganzfeld* procedure [Bem & Honorton, 1994; Hyman, 1994].)

Thus even in the rare cases where the literal truth of the null hypothesis is at issue, and especially in the preponderance of cases where the null hypothesis is a straw man, the investigator wants to formulate a position going beyond a primitive test of a single null hypothesis. Scientific arguments are much more rich than that. Before working up to the issues in some typical debates, however, we stay at an elementary level and discuss in detail what makes statistical arguments rhetorically forceful and narratively compelling.

PERSUASIVE ARGUMENTS: THE MAGIC CRITERIA

There are several properties of data, and its analysis and presentation, that govern its persuasive force. We label these by the acronym MAGIC, which stands for *magnitude*, *articulation*, *generality*, *interestingness*, and *credibility*.[5]

[5]There are other schemes for classifying the quality of statistical evidence and its presentation. The enumeration of various forms of *validity* (internal, external, construct, trait, discriminant, ecological, predictive, etc.) is a well-known alternative (Campbell,

Magnitude

The strength of a statistical argument is enhanced in accord with the quantitative magnitude of support for its qualitative claim. There are different ways to index magnitude, the most popular of which is the so-called "effect size" (Cohen, 1988; Glass, 1978; Hedges & Olkin, 1985; Mullen, 1989; Rosenthal, 1991). In the basic case of the comparison between two means, effect size can be simply given as the difference between the means; often, however, this difference is divided by the standard deviation of observations within groups. In chapter 3, we bring up a number of alternatives, and introduce the concept of "cause size," which also bears on the interpretation of magnitudes of effects.

Articulation

By articulation, we refer to the degree of comprehensible detail in which conclusions are phrased. Suppose, for example, that the investigator is comparing the mean outcomes of five groups: A, B, C, D, E. The conclusion "there exist some systematic differences among these means" has a very minimum of articulation. A statement such as, "means C, D, and E are each systematically higher than means A and B, although they are not reliably different from each other" contains more articulation. Still more would attach to a quantitative or near-quantitative specification of a pattern among the means, for example, "in moving from Group A to B to C to D to E, there is a steady increase in the respective means." The criterion of articulation is more formally treated in chapter 6, where we introduce the concepts of *ticks* and *buts*, units of articulation of detail.

Generality

Generality denotes the breadth of applicability of the conclusions. The circumstances associated with any given study are usually quite narrow, even though investigators typically intend their arguments to apply more broadly. To support broad conclusions, it is necessary to include a wide range of contextual variations in a comprehensive research plan, or to cumulate outcome data from many interrelated but somewhat

1960; Cook & Campbell, 1979). The analysis of validity has been very useful, but it has never caught on as a coherent whole. It strikes the student as rather formal and esoteric. Another system has been given in an exquisitely sensible little book on the importance of statistical analysis in medical research (Hill, 1977). This author named but did not elaborate on criteria similar to mine. If my approach has any claim to novelty, it is that I have chosen and developed my criteria within the unifying premise of statistics as argument, and I know of no previous source that has systematically pursued such an approach.

different studies, as can be done within the context of meta-analysis (Mullen, 1989; Rosenthal, 1991). In chapter 7, we present an analysis of variance framework for interpreting generalization.

High-quality evidence, embodying sizeable, well-articulated, and general effects, is necessary for a statistical argument to have maximal persuasive impact, but it is not sufficient. Also vital are the attributes of the research story embodying the argument. We discuss two criteria for an effective research narrative: interestingness, and credibility.

Interestingness

Philosophers, psychologists, and others have pondered variously what it means for a story to be interesting (e.g., Davis, 1971; Hidi & Baird, 1986; Schank, 1979; Tesser, 1990), or to have a point (Wilensky, 1983). Our view in this book is that for a statistical story to be *theoretically* interesting, it must have the potential, through empirical analysis, to change what people believe about an important issue. This conceptual interpretation of statistical interestingness has several features requiring further explanation, which we undertake in chapter 8. For now, the key ideas are *change of belief*—which typically entails surprising results—and the *importance* of the issue, which is a function of the number of theoretical and applied propositions needing modification in light of the new results.

Credibility

Credibility refers to the believability of a research claim. It requires both *methodological* soundness, and *theoretical* coherence. Claims based on sloppy experimental procedures or mistaken statistical analyses will fall victim to criticism by those with an interest in the results. Clues suggested by funny-looking data or wrongly framed procedures provide skeptics with information that something is amiss in the statistical analysis or research methodology. (Of course, you might yourself be in the role of critic, whereby you can track these clues in other people's research reports.)

Many extensive catalogs of methodological and statistical errors already exist in the literature (Aronson, Brewer, & Carlsmith, 1985; Campbell & Stanley, 1963; Evans, 1991; King, 1986). Our discussions differ from standard ones in two ways. We classify statistical errors "bottom up"—that is, in terms of various odd appearances in data, from which types of error may be induced (chap. 5); and we treat a selection of research design errors in the context of how they might affect the ongoing debate between an investigator and a critic (chap. 9).

The credibility of a research claim can sustain damage from another source—the claim may violate prevailing theory, or even common sense. The research audience cannot bring itself to believe a discrepant claim, such as a purported demonstration of extrasensory perception, which would require vast revision of existing views. In such cases, debate tends to occur on two fronts simultaneously. The critic will typically pick on suspected methodological errors, thus accounting for the claim as a methodological artifact. The investigator must be prepared to try to rule out such accounts. Also, a theoretical battle will develop, in which the investigator is challenged to show that her alternative theory is *coherent*, that is, capable of explaining a range of interconnected findings. If result A requires explanation X, result B calls forth explanation Y, and result C explanation Z, where explanations X, Y, and Z have little relation to each other, the narrative of results A, B, and C is incoherent (Thagard, 1989). On the other hand, if a single explanatory principle accounts for several different results, the story is coherent. When the results would be unrelated were it not for sharing the same explanation, the story is not only coherent, it is *elegant*. In chapter 9, we refer to coherent bundles of results as *signatures*.

The outcome of a theoretical debate depends on the comparative adequacy of the respective accounts of existing data. But the contest may also hinge on who has the *burden of proof* in the exchange of criticisms and rebuttals. Usually, this burden rests with the investigator, especially at the outset. Critics are often freewheeling in their invention of counterexplanations: It could be this, it may be that, it's merely such-and-so. Some types of counterexplanations are so vague as to be untestable—which gives the critic a substantial debating advantage. Nevertheless, despite occasional abuse of the ability to criticize, science is better off being tolerant of kibitzers and second-guessers. The critic is often right. Anyway, science should have both a conservative bias—which prevents rapid and bewildering shifts of views—and ultimate openness, such that persistent innovators can ultimately triumph if their claims are indeed meritorious. These issues are discussed more deeply in chapter 9.

The requisite skills for producing credible statistical narratives are not unlike those of a good detective (Tukey, 1969). The investigator must solve an interesting case, similar to the "whodunit" of a traditional murder mystery, except that it is a "howcummit"—how come the data fall in a particular pattern. She must be able to rule out alternatives, and be prepared to match wits with supercilious competitive colleagues who stubbornly cling to presumably false alternative accounts, based on somewhat different clues. (This is analogous to the problems faced by heroic fictional detectives who must put up with interference from cloddish police chiefs.)

STYLE AND CONVENTION

Our five major criteria for an effective statistical argument depend on the quality of the data and on the skill of the investigator in designing research and presenting results. There are other aspects of statistical arguments that depend hardly at all on data or on skill—instead, they are matters of taste and convention.

Style

For our purposes, the *style* of the statistical argument within a research presentation can be loosely represented by a dimension along which different possible presentations of the same results can be arrayed: At one extreme is an assertive and incautious style, running toward reckless and excessive claims; at the other extreme is a timid and rigid style, with an unwillingness to make any claims other than the most obvious ones. In practice, styles are not usually at the extremes, but rather at intermediate positions nearer to one pole than the other. We label these the *liberal* style and the *conservative* style (chap. 4).

The liberal style is oriented more toward exploration of data and discovery of possibly systematic effects. By contrast, the conservative style reflects a confirmatory attitude toward research results, where one is willing to forego claims about marginal or unexpected findings in order to be more confident about the remaining claims.

It might seem that one should be able to calibrate just how liberal to be if one could place relative costs on making too many claims versus making too few claims. Indeed, there are times when research is explicitly exploratory, with open season on speculations, and times when it is explicitly confirmatory, requiring the utmost prudence. But most research falls somewhere in the middle, and even in explicit cases, the required decision calculation is impractical because the costs of the two types of errors are not sensibly quantifiable. There is a boundary in data interpretation beyond which formulas and quantitative decision procedures do not go, where judgment and style enter.

Conventions

Sometimes this subjective element is disguised by the use of *conventions*. There are many such, the most prominent of which is the notorious $p = .05$ as the "conventional" significance level. If everyone follows the conventions, individual investigators are freed from the responsibility (but denied the opportunity) for using their own judgment. This is relatively benign as long as conventions are reasonable, and everyone realizes that they are conventions rather than commandments.

The Inevitability of Uncertainty

An analogy can be drawn with a legal system. The dispensation of justice is fraught with uncertainty. There are imponderable costs associated with declaring a guilty person innocent, or an innocent person guilty. The balance between these two types of mistake is set by the legal conventions of a society, in particular, how weighty the evidence of a defendant's guilt must be to justify convicting him. In the Anglo-American tradition for capital offenses, guilt must be established "beyond a reasonable doubt." Such a convention, though it may convey a reassuring illusion that the decision policy is exact (provided that nobody is lying), is itself subject to ambiguity and alternative interpretation. By and large, nonetheless, wise use of this imprecise tradition serves us well.

In applications of statistics in the social sciences, some element of subjectivity is always present, and the research investigator is cast in a role analogous to that of a legal advocate. In this metaphor, the scientific audience plays the part of judge or jury hearing the testimony of the investigator and of those who may disagree. Though it may take several judicial proceedings, a judgment is eventually reached.

THE BOTTOM LINE

A research story can be interesting and theoretically coherent, but still not be persuasive—if the data provide only weak support for the rhetoric of the case. On the other hand, a lot of high-quality rhetoric can be squandered by a poor narrative—for example, if the research is so dull that no one cares which way the results come out. Thus rhetoric and narrative combine multiplicatively, as it were, in the service of persuasive arguments based on data analysis. If either component is weak, the product is weak. The argument is strong only when it has the MAGIC properties of forceful rhetoric and effective narrative. In making his or her best case, the investigator must combine the skills of an honest lawyer, a good detective, and a good storyteller.

2 Elementary Arguments and the Role of Chance

I have proposed that the proper function of statistics is to formulate good arguments explaining comparative differences, hopefully in an interesting way. In this chapter the four most elementary kinds of statistical arguments are introduced, with a pithy illustration of each. All four are affected by some manifestation of chance. The first two arguments each involve a comparison between a set of observations and a chance expectation; the arguments differ in whether or not a comparative difference is claimed. The next pair of arguments compare the means of two sets of observations with each other; one argument claims that the observed mean difference could have been due to chance, whereas the other claims a systematic difference on top of chance effects.

Common sense tells us that an explanatory argument is better if concise, with overexplanation giving grounds for skepticism. If John excuses his absence from work by claiming that his grandmother died, and besides, he had a bad cold, we are inclined to doubt his sincerity. He has violated one of Grice's (1975) axioms of normal discourse, namely that a speaker should give sufficient information to be understood, but no more.

As with everyday explanations, even more so with scientific explanations. The ideal of parsimony gives preference to the simplest explanation adequate to account for a given corpus of data. In chapter 1, we noted that in the social sciences, part of the job of the investigator is to sort out the relative contributions to data from chance factors and from systematic factors. The most parsimonious (though often not the most welcome) kind of explanation is that the data can be attributed entirely to chance factors. This case of pure chance serves as a very low-level, baseline explanation that may be altered to include systematic factors if statistical analysis shows them to be necessary. The role of chance in the four simple arguments (and more complex ones to come later) is not entirely transparent, and requires discussion. In this chapter we offer two conceptions of chance—random *generation* and random *sampling*—